OpenAI

1960 Bryant Street
San Francisco, CA 94110

June 12, 2023

From:
Anna Makanju
VP of Public Policy
OpenAI

To:
Stephanie Weiner
Acting Chief Counsel
NTIA

OpenAI is pleased to respond to the National Telecommunications and Information Administration's (NTIA)'s April 13, 2023 request for comments (RFC) on AI Accountability Policy.

In this comment, we describe our thinking on AI accountability based on the safety practices we apply to the services we provide today, and the practices we plan to apply to services we anticipate offering in the future. We welcome NTIA's decision to frame this discussion in terms of an "ecosystem" of AI accountability. As the RFC observes, policy stakeholders are exploring a "range of trustworthy AI system goals and deployment contexts."[1] Policies and practices designed to achieve accountability will vary accordingly. At the same time, specific accountability measures will need to coexist with one another, and what matters most is the impact they have in concert.

We believe that a mature ecosystem for AI accountability will include both horizontal and vertical elements. That is, we both expect there to be some elements that apply to certain AI systems across domains of application, as well as some elements that are tailored to particular domains. We are engaged in the development and deployment of highly capable foundation models – models that learn from a large amount of data in order to be able to perform a wide range of downstream tasks. In our view, AI developers like us must act responsibly and take a careful and safety-focused approach to the development and deployment of the most advanced capabilities. This is true regardless of the particular domains in which such models may be used.

A wide range of existing laws already apply to AI – including to our products – and the legal landscape is quickly evolving, with legislative initiatives in Congress, the AI Act under development in Europe, and legislative and policy initiatives unfolding around the world. At the same time, long-established bodies of law, regulation, and other expectations in areas like medicine, education, and employment are already being interpreted and adapted in ways that will shape the role AI plays in those domains. We see these sector-specific efforts, informed by deep domain expertise, as a critical part of the AI accountability landscape.

We strongly support efforts to harmonize the emergent accountability expectations for AI, including the efforts of the NIST AI Risk Management Framework, the U.S.-E.U. Trade and Technology Council, and a range of

[1] https://ntia.gov/issues/artificial-intelligence/request-for-comments

other global initiatives. While these efforts continue to progress, and even before new laws are fully implemented, we see a role for ourselves and other companies to make voluntary commitments on issues such as pre-deployment testing, content provenance, and trust and safety.

Our current engineering approach requires a unique scale of computing resources, and we regard this as a promising basis for defining additional and distinctive accountability expectations that would apply to actors like us. We support scoping any new regulation for highly capable foundation models carefully so as to preserve the ability of all actors to fairly compete and innovate.

Accountability plays a role throughout the technology lifecycle. Our efforts to make our models safe and reliable begin before development starts, continue throughout deployment and operation of our models, and address both creators and users of highly capable foundation models. We provide developers with world-leading capabilities for their applications, and provide powerful capabilities directly to the millions of people who use ChatGPT and our other services every day. Our usage policies apply to all users of our models, tools, and services.[2] We comply with existing laws, and require that our developers and users comply when they use our services.

We focus the remainder of this comment on our current approaches to AI accountability, and describe important areas where we and others are working to strengthen the ecosystem. We note that policymakers in the United States and around the world are considering a wide range of policies and measures intended to achieve AI accountability, including legislation, regulations, international agreements, self-regulatory programs, and enforceable technical and other standards. We appreciate these efforts and stand ready to partner with other stakeholders to develop and implement effective approaches to AI accountability.

OpenAI's Current Approaches

We are refining our practices in tandem with the evolving broader public conversation. Here we provide details on several aspects of our approach.

System Cards

Transparency is an important element of building accountable AI systems. A key part of our approach to accountability is publishing a document that we currently call a System Card, for new AI systems that we deploy. Our approach draws inspiration from previous research work on model cards

[2] https://openai.com/policies/usage-policies

and system cards.[3] To date, OpenAI has published two system cards: the GPT-4 System Card and DALL-E 2 System Card.[4]

We believe that in most cases, it is important for these documents to analyze and describe the impacts of a system – rather than focusing solely on the model itself – because a system's impacts depend in part on factors other than the model, including use case, context, and real world interactions. Likewise, an AI system's impacts depend on risk mitigations such as use policies, access controls, and monitoring for abuse. We believe it is reasonable for external stakeholders to expect information on these topics, and to have the opportunity to understand our approach.

Our System Cards aim to inform readers about key factors impacting the system's behavior, especially in areas pertinent for responsible usage. We have found that the value of System Cards and similar documents stems not only from the overview of model performance issues they provide, but also from the illustrative examples they offer. Such examples can give users and developers a more grounded understanding of the described system's performance and risks, and of the steps we take to mitigate those risks. Preparation of these documents also helps shape our internal practices, and illustrates those practices for others seeking ways to operationalize responsible approaches to AI.

Qualitative Model Evaluations via Red Teaming

Red teaming is the process of qualitatively testing our models and systems in a variety of domains to create a more holistic view of the safety profile of our models. We conduct red-teaming internally with our own staff as part of model development, as well as with people who operate independently of the team that builds the system being tested. In addition to probing our organization's capabilities and resilience to attacks, red teams also use stress testing and boundary testing methods, which focus on surfacing edge cases and other potential failure modes with potential to cause harm.

Red teaming is complementary to automated, quantitative evaluations of model capabilities and risks that we also conduct, which we describe in the next section. It can shed light on risks that are not yet quantifiable, or those for which more standardized evaluations have not yet been developed. Our prior work on red teaming is described in the DALL-E 2 System Card and the GPT-4 System Card.

Our red teaming and testing is generally conducted during the development phase of a new model or system. Separately from our own

[3] https://arxiv.org/abs/1810.03993, https://montrealethics.ai/system-cards-for-ai-based-decision-making-for-public-policy/
[4] https://cdn.openai.com/papers/gpt-4-system-card.pdf, https://github.com/openai/dalle-2-preview/blob/main/system-card.md

internal testing, we recruit testers outside of OpenAI and provide them with early access to a system that is under development. Testers are selected by OpenAI based on prior work in the domains of interest (research or practical expertise), and have tended to be a combination of academic researchers and industry professionals (e.g, people with work experience in Trust & Safety settings). We evaluate and validate results of these tests, and take steps to make adjustments and deploy mitigations where appropriate.

OpenAI continues to take steps to improve the quality, diversity, and experience of external testers for ongoing and future assessments.

Quantitative Model Evaluations

In addition to the qualitative red teaming described above, we create automated, quantitative evaluations for various capabilities and safety oriented risks, including risks that we find via methods like red teaming. These evaluations allow us to compare different versions of our models with each other, iterate on research methodologies that improve safety, and ultimately act as an input into decision-making about which model versions we choose to deploy. Existing evaluations span topics such as erotic content, hateful content, and content related to self-harm among others, and measure the propensity of the models to generate such content.

Usage Policies

OpenAI disallows the use of our models and tools for certain activities and content, as outlined in our usage policies.[5] These policies are designed to prohibit the use of our models and tools in ways that cause individual or societal harm. We update these policies in response to new risks and updated information about how our models are being used. Access to and use of our models are also subject to OpenAI's Terms of Use which, among other things, prohibit the use of our services to harm people's rights, and prohibit presenting output from our services as being human-generated when it was not.[6]

We take steps to limit the use of our models for harmful activities by teaching models to refuse to respond to certain types of requests that may lead to potentially harmful responses. In addition, we use a mix of reviewers and automated systems to identify and take action against misuse of our models. Our automated systems include a suite of machine learning and rule-based classifier detections designed to identify content that might violate our policies. When a user repeatedly prompts our models with policy-violating content, we take actions such as issuing a

[5] https://platform.openai.com/docs/usage-policies/use-case-policy
[6] https://openai.com/policies/terms-of-use

1960 Bryant Street
San Francisco, CA 94110

warning, temporarily suspending the user, or in severe cases, banning the user.

Open Challenges in AI Accountability

As discussed in the RFC, there are many important questions related to AI Accountability that are not yet resolved. In the sections that follow, we provide additional perspective on several of these questions.

Assessing Potentially Dangerous Capabilities

Highly capable foundation models have both beneficial capabilities, as well as the potential to cause harm. As the capabilities of these models get more advanced, so do the scale and severity of the risks they may pose, particularly if under direction from a malicious actor or if the model is not properly aligned with human values.

Rigorously measuring advances in potentially dangerous capabilities is essential for effectively assessing and managing risk. We are addressing this by exploring and building evaluations for potentially dangerous capabilities that range from simple, scalable, and automated tools to bespoke, intensive evaluations performed by human experts. We are collaborating with academic and industry experts, and ultimately aim to contribute to the development of a diverse suite of evaluations that can contribute to the formation of best practices for assessing emerging risks in highly capable foundation models. We believe dangerous capability evaluations are an increasingly important building block for accountability and governance in frontier AI development.

Open Questions About Independent Assessments

Independent assessments of models and systems, including by third parties, may be increasingly valuable as model capabilities continue to increase. Such assessments can strengthen accountability and transparency about the behaviors and risks of AI systems.

Some forms of assessment can occur within a single organization, such as when a team assesses its own work or when a team or part of the organization produces a model and another team or part, acting independently, tests that model. A different approach is to have an external third party conduct an assessment. As described above, we currently rely on a mixture of internal and external evaluations of our models.

Third-party assessments may focus on specific deployments, a model or system at some moment in time, organizational governance and risk

1960 Bryant Street
San Francisco, CA 94110

management practices, specific applications of a model or system, or some combination thereof. The thinking and potential frameworks to be used in such assessments continue to evolve rapidly, and we are monitoring and considering our own approach to assessments.

For any third-party assessment, the process of selecting auditors/assessors with appropriate expertise and incentive structures would benefit from further clarity. In addition, selecting the appropriate expectations against which to assess organizations or models is an open area of exploration that will require inputs from different stakeholders. Finally, it will be important for assessments to consider how systems might evolve over time and build that into the process of an assessment / audit.

Registration and Licensing for Highly Capable Foundation Models

We support the development of registration and licensing requirements for future generations of the most highly capable foundation models. Such models may have sufficiently dangerous capabilities to pose significant risks to public safety; if they do, we believe they should be subject to commensurate accountability requirements.

It could be appropriate to consider disclosure and registration expectations for training processes that are expected to produce highly capable foundation models. Such disclosure could help enable policymakers with the necessary visibility to design effective regulatory solutions, and get ahead of trends at the frontier of AI progress. It is crucial that any such regimes prioritize the security of the information disclosed.

AI developers could be required to receive a license to create highly capable foundation models which are likely to prove more capable than models previously shown to be safe. Licensure is common in safety-critical and other high-risk contexts, such as air travel, power generation, drug manufacturing, and banking. Licensees could be required to perform pre-deployment risk assessments and adopt state-of-the-art security and deployment safeguards; indeed, many of the accountability practices that the NTIA will be considering could be appropriate licensure requirements. Introducing licensure requirements at the computing provider level could also be a powerful complementary tool for enforcement.

There remain many open questions in the design of registration and licensing mechanisms for achieving accountability at the frontier of AI development. We look forward to collaborating with policymakers in addressing these questions.